(ICW) Meta Leaks Part 2: How To Kill A Social Movement

2025-08-14

By NRU

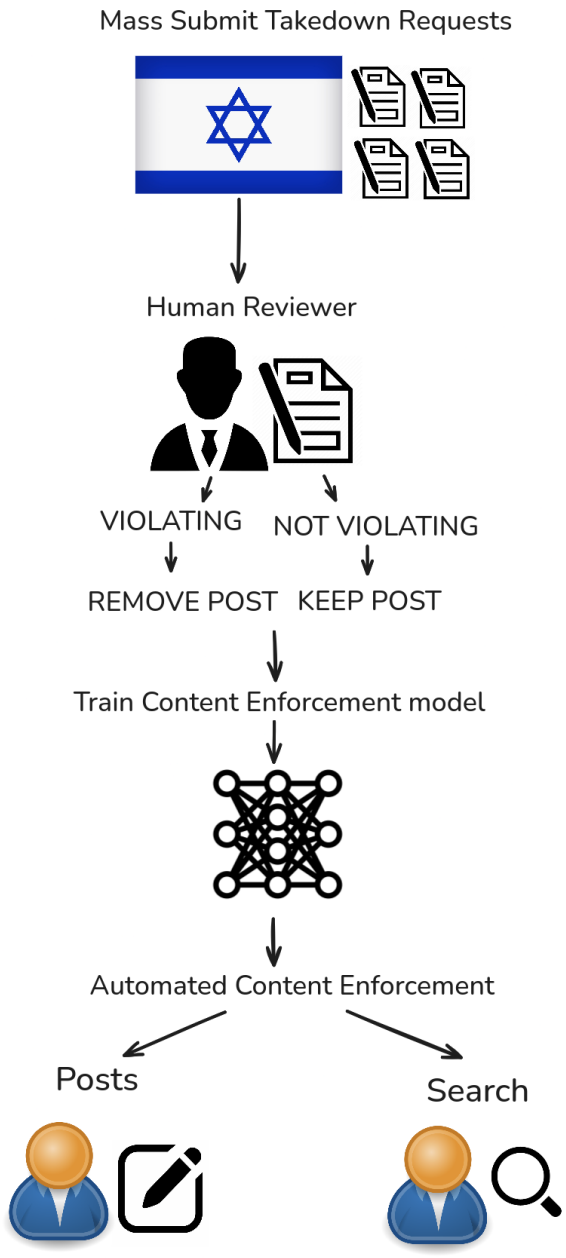

Figure 0. Example of how the Israeli government uses mass takedown requests to takedown posts and reduce the discoverability of certain information in search

International Corruption Watch

Perilous times are coming

The International Corruption Watch (ICW) is a newly founded organization of independent journalists with the goal of bring to light the massive corruption that plagues governments and big tech worldwide. In an era where technology plays an increasingly central role in shaping public policy and opinion, we believe that the misuse of technology for corrupt purposes has become one of the most pressing issues of our time.

Follow us on social media:

BlueSky: https://bsky.app/profile/icw-nru.bsky.social

Substack: https://substack.com/@icwnru

Do you have any tips for a case of corruption in tech or government?

If so please send a message to **icw_nru@protonmail.com**

ICW is a self-funded operation. All donations are appreciated.

BTC: bc1qjy84xyr8u0majvtslzfz940hpkcp4hv3gjmms8

ETH: 0x0620C26611644Ec7287274713df0eD3Df93dBB82

Read our previous investigations!

(2024-11-17) The Youtube Alogirthm and Manufacturing Consent

(2024-11-30) What can we learn from the Andrew Tate data breach?

(2025-08-11) (ICW) Meta Leaks Part 1: Israel & Meta, The Greatest Global Mass Campaign to Ever Exist

Note from NRU

The Meta leaks series aim to expose corruption inside Meta. Every piece of information and accusation is brought forth by Meta employees. The whistleblowers' motivation behind these leaks is the following: Stop all involvement with the Israeli government and their current genocide in Gaza. Until then, more leaks will be dropped. With each leak exposing a different aspect of corruption related to censorship, AI, and financial crimes.

Since the first leak, Meta has yet to take any step in mitigating the exploitation of its platform.

Summary

We have obtained internal leaks from Meta employees in the “Integrity” organization. The results of our investigation suggest that:

- Meta’s platforms have been used to effectively censor posts, and reduce discoverability of social movements by removing related hashtags and topics from search results.
- We present three recent examples of this: hiding search results related to the American Democratic party, hiding search results that are explicitly anti-Israel, and suppressing posts expressing dissent against the government of Bangladesh.
    - Meta hid search results related to the American Democratic party such as posts containing hashtags #democrats, #voteblue, and #bluewave, while the results related to the American Republican party were unharmed. The over-enforcement of Democrat-adjacent hashtags took place since November 2024, the same month as the American election.
    - Hashtags that are anti-Israel such as #israelicrimes and #sanctionisraelnow were significantly more likely to get censored than their pro-Israel equivalents
    - The Bangladeshi government is responsible for the sixth highest number of takedown requests on Meta’s platform. Bangladesh's takedown requests coincide heavily with days of social unrest and protests, such as the student protests in 2024.
- We also put forth the following claim against Meta. Meta is likely complicit in the mass censorship of posts and information that are orchestrated by bad actors.
    - Meta is aware of the fact that their content enforcement models are flawed
    - Meta has not taken any concrete steps to provide long-term fixes to their system
    - Any fix done by Meta so far has been short-term, for example, whitelisting false-positives
    - Meta is aware that bad actors such as the Israeli government continue to abuse their system and have yet to do anything about it so far.

Meta leaks part 1 recap – How content enforcement works

Training

In order to have an AI that can moderate content, Meta first needs to train a machine learning (ML) model. The "How enforcement technology works" page partially describes the process:

1. Review teams review a post to assess whether it violates the community guidelines.
2. Reviewer decisions will be used as a training data for Meta's content enforcement machine learning (ML) models. In other words, the ML model learns from human reviewer data, so that it can be used for the content review and enforcement process at scale.

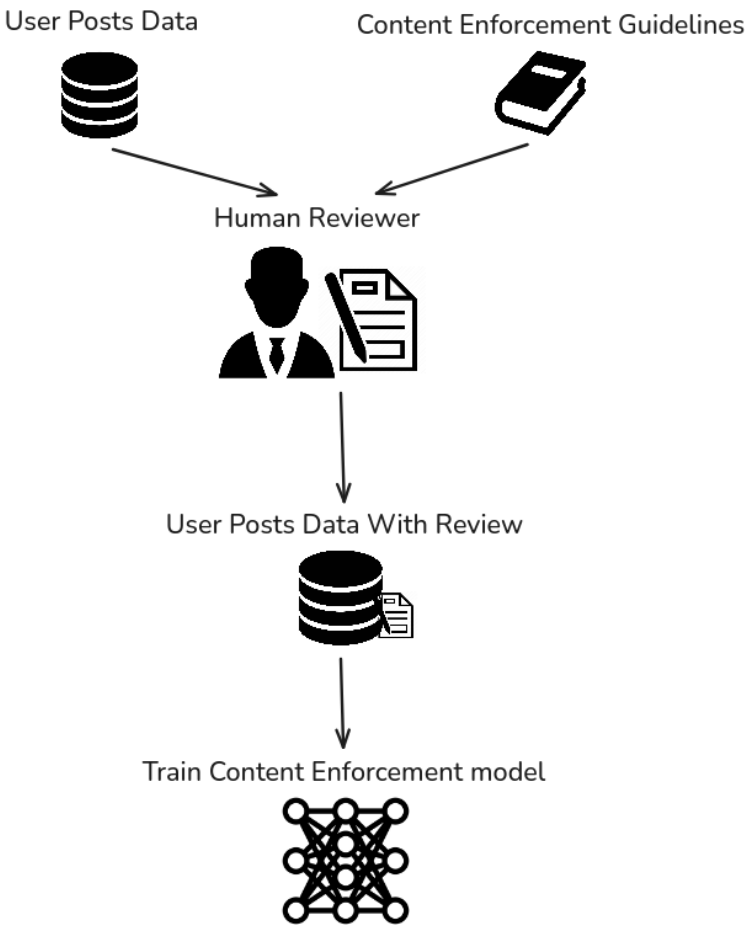

Figure 1. How data is reviewed and used for training at Meta's Integrity organization

Now that the model is trained, it is ready to be used in production (live traffic on Facebook and Instagram). Meta will now use the model to classify every new post. According to Meta, 90% of the posts on its platforms are actioned upon before anyone reports it. This is done using the content enforcement ML models. The "How enforcement technology works" page partially describes the process:

- Once a post gets added to the platform, the content enforcement model will classify the post as VIOLATING or NOT VIOLATING, along with a confidence score.
- If a high confidence score is given and the post IS VIOLATING, it gets removed. If NOT VIOLATING, it will stay.
- If however a low confidence score is given, a human reviewer will review the post and provide the final enforcement decision themselves.

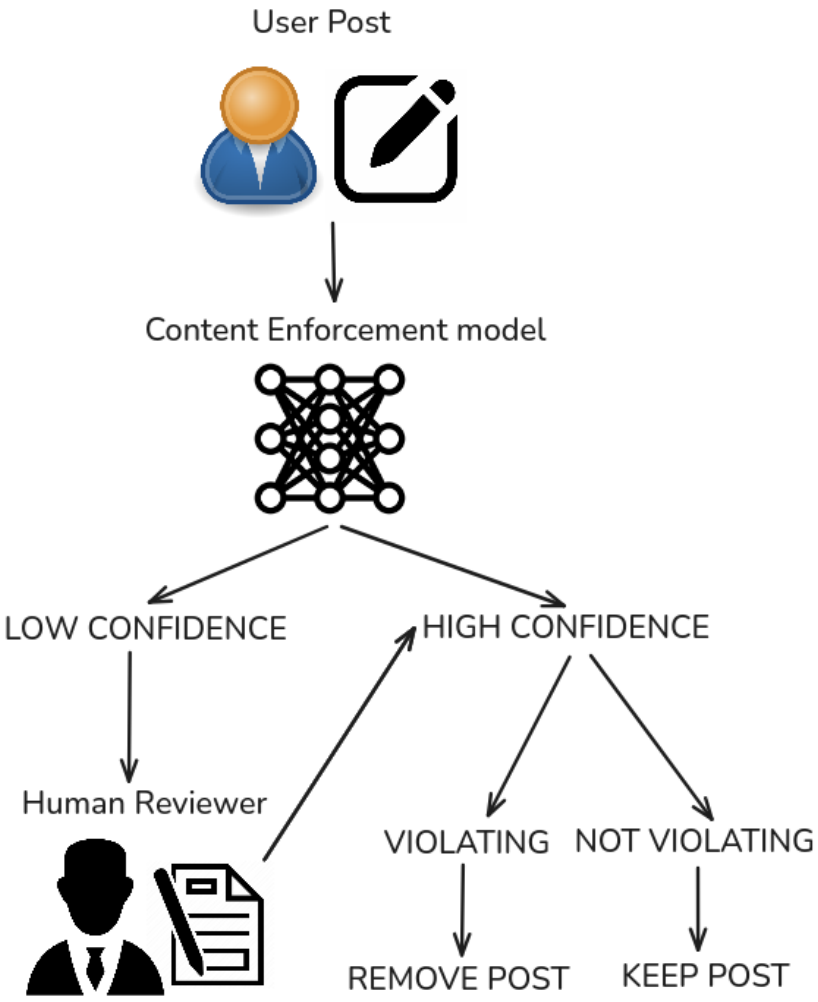

Figure 2. How the enforcement model is used to classify all incoming user posts on Meta's platform

How does search recommendation work at Meta?

The leaks from part 1 are quite relevant to this story. In part 1, I showed how content enforcement models can easily be abused by poisoning their training dataset. Once a model is poisoned, it will censor virtually all new posts made on Meta's platforms.

However, content enforcement models play an important role at Meta beyond post moderation. I am specifically referring to information discoverability.

On Meta's platforms, there are three main entry-points for information access:

1. Recommended posts in user feed
2. Search auto-complete
3. Search results

The 'Meta leaks part 1' covered the first entry-point, user posts. This investigation will focus on the remaining two, search related information discovery.

Search auto-complete, known internally as TypeAhead, are the recommendations the user gets when they begin to type something in the search bar.

facebook
how to save on power
Search the Web
how to save on power bill
how to save on power
how to save on power point
how to save on powder toy
how to save on powder game
See more results for how to save on power
Displaying top 5 results

Figure 3. Search auto-complete example on facebook.com

Search results, known internally as SERP (search engine results page), are the recommendations the user gets after they enter a term in the search bar.

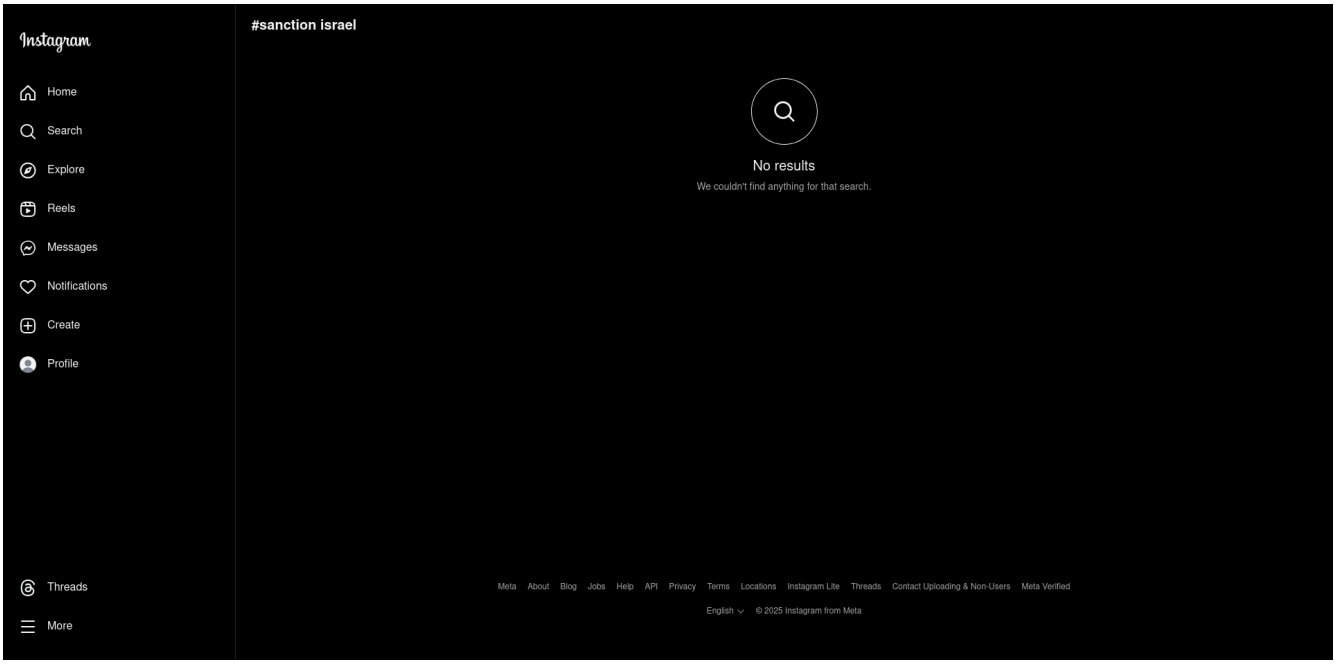

Figure 4. Search results for #sanctionisrael on instagram.com. Note that no results appear for the searched term.

At Meta, as well as other content based platforms, content is served through a recommendation engine. AI powered recommendation engines have become standard across all these platforms including Meta.

I have covered how AI powered recommendation engines work in my previous investigation titled "The YouTube Algorithm and Manufacturing Consent".

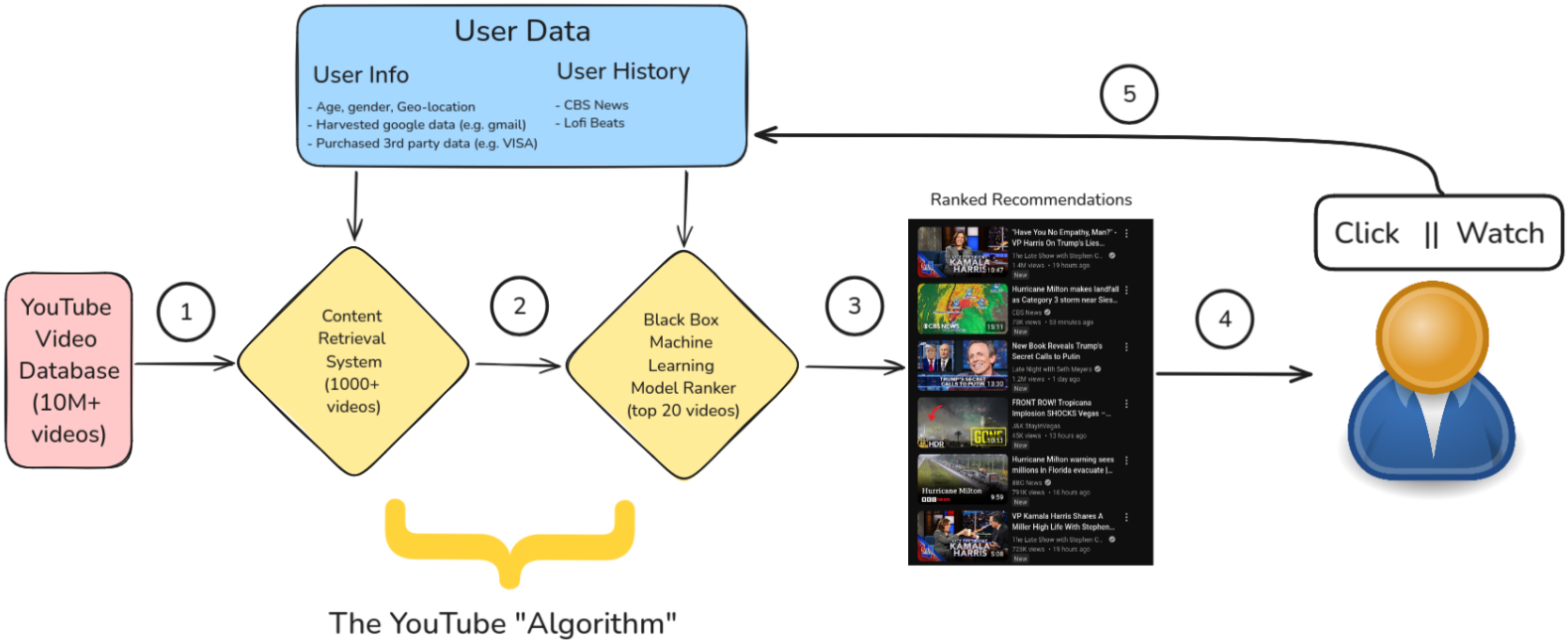

Figure 5. Simplified version of YouTube's recommendation system, aka, the YouTube algorithm.

Meta’s content recommendation systems for Facebook and Instagram are no different. Almost every recommendation on Meta’s platform is powered by AI and trained on user data.

How do content enforcement models affect search?

On most social media platforms, users aren't always recommended the final outputs of a recommendation system. At Meta, content enforcement models act as one of the final gatekeepers for the content delivered to the end user.

In the case of user posts, the posts would outright get removed or down-ranked. In the case of search, the results would be omitted entirely.

So why does a hashtag's results get removed from search? Meta has an automated hashtag removal system. If a hashtag is found to contain a certain number of violating posts on any given day, the hashtag's search results will automatically get removed. The content enforcement models determine whether the post is violating, and the number of violating posts required is likely set by an internal threshold parameter.

From the outside, this may seem quite confusing. Users expect that hashtags that are inherently violating to be the ones that are removed (e.g. hashtag inciting hatred such as 'x_ethnicity_is_scum'). Its quite unusual for the visibility of a hashtag or topic to be contingent on the posts associated with them at a specific point in time.

With the current system, topics like 'x_ethnicity_is_scum' and 'united_nations' are treated equally. They can both be visible and recommended one day, or completely removed in another day.

How to kill a social movement

If you are a bad actor who really wants to reduce the reach of a social movement on Meta's platforms, then doing so is quite simple. Here is a step by step guide on how to do it.

1. Find a social movement you want to take down. Find posts and hashtags associated with it.
2. Hijack the hashtag by posting content that is likely to get flagged by Meta's content enforcement model. Almost anything related to Palestine/Gaza will likely be flagged right away.
3. Produce mass takedown requests targeting the type of posts you want gone. Submitting them through government specific channels increases the success of a takedown exponentially (93% in case of the Israeli government). If Meta doesn't comply, threaten them with advertiser boycotts, fines, or nation-wide bans.

At this point, all the bad actor has to do is keep periodically producing violating posts under a hashtag and sending in takedown requests. Once the above is done, Meta will kill off the social movement by:

1. Removing posts related to the ones found in the takedown requests

2. Remove search for hashtags which contain content that the content enforcement model flags
3. As discoverability for the content is significantly reduced and users are given strikes or banned, user engagement will naturally drop. Since users are no longer engaging with this type of content, Meta's recommendation systems will be less likely to recommend similar types of content in the future. This is essentially a negative feedback loop.

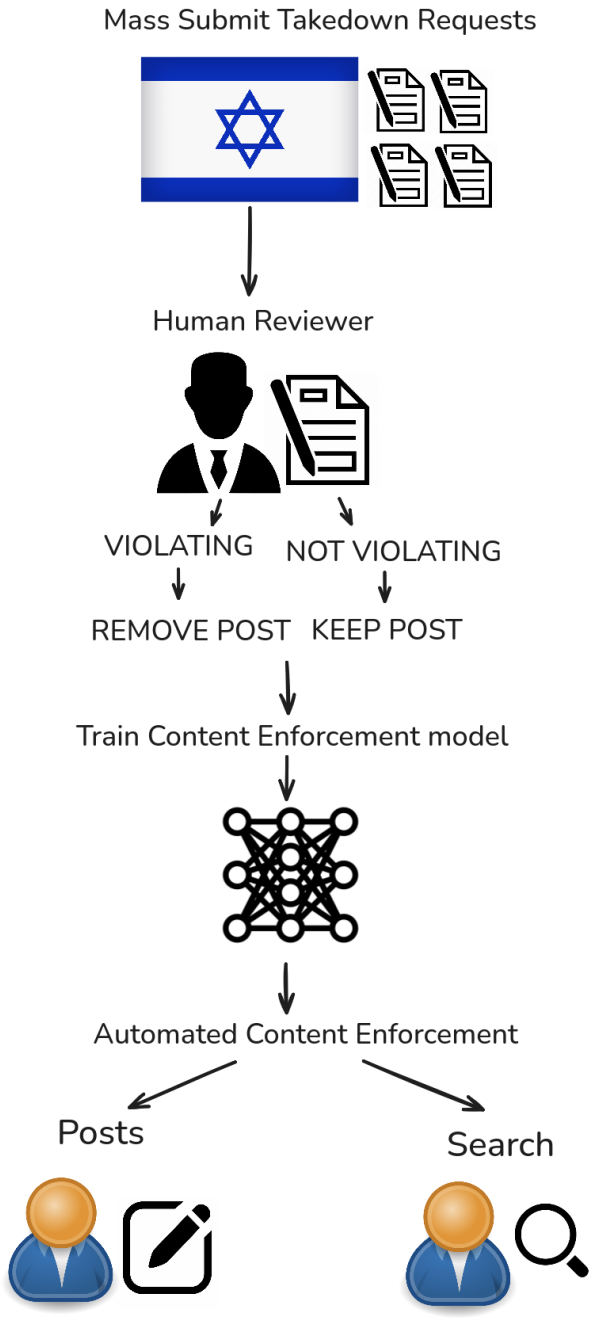

Figure 6. Example of how the Israeli government uses mass takedown requests to takedown posts and reduce the discoverability of certain information in search

Real world examples, real consequences

In this section, I will present three recent examples where Meta's content enforcement models were exploited by bad actors with full knowledge and complicity of Meta in order to facilitate platform-wide censorship.

The three cases are:

1. Suppressing search results associated with the American Democratic party
2. Censoring posts and suppressing search results that are explicitly anti-Israel
3. Censoring posts that express dissent against government actions in Bangladesh

Case #1: Democrats

Remember the story from January 2025, where Meta was accused of censoring Democrat aligned posts? The reason for the removal of search results for 'Democrats' was not because Meta was deliberately placing a filter on them (as opposed to the time where they increase their filter thresholds for posts related to Gaza in the Middle East region).

Meta has actually been over-enforcing Democrat related hashtags since November 2024. The issue only became visible in January 2025, due to an internal bug, which made the over-enforcement issue much worse than before.

Internally, the reason why Democrat related hashtags were censored was because of false-positive signals from the content enforcement models. The content enforcement model classified content under democrat-adjacent hashtags as mostly misinformation and sometimes terrorism.

Furthermore, Meta discovered a large surge in user reports for posts under democrat-adjacent hashtags, which further fueled the over-enforcement of the models. It is still unknown who or what group is responsible for these mass reports.

One thing to mention is that Meta doesn't actively monitor censored hashtags, even if they are associated with important topics such as elections. After the Democrat scandal, almost all terms related to the American Democratic party were whitelisted. For other countries, they are out of luck. It wouldn't be surprising if this system was and continues to be exploited by bad actors, particularly governments.

In the Meta leaks part 1, I showed how the terrorism category is almost always related Israel-Palestine content. Which brings the question why are certain hashtags associated with the Democratic party being flagged for terrorism? Is Israel-Palestine content being posted under Democrat-adjacent hashtags? Was it done deliberately by someone familiar with the content enforcement system? And by whom?

Case 2: Israel

After October 7 2023, Meta has not only been complicit and enabled mass censorship of pro-Palestinian and anti-Israel posts, but has also decreased discoverability of such content as well.

Please read ‘The Meta leaks part 1” to learn about how the posts were censored. This investigation will focus more on search and information discoverability.

Due to the overly-biased content enforcement model, posts that are anti-Israel are likely to get flagged by these models, and as a result, the hashtags for which they fall under are also likely to get removed.

Below is a small list of the hashtags that are known to have been censored on Meta’s platform. This list is pulled directly from internal tracking documents written in late 2023 at Meta’s integrity organization. In other words, Meta knows that these terms are blocked, and has done nothing about it so far.

#fuckisrael

#israeliterrorists

#israelicrimes

#israelkillchildren

#israelinewnazism

#israel_is_killing_innocent_people_in_gaza

#sanctionisraelnow

#israel_hamas_are_murders

#istandwithhamas

Out of all of these hashtags, only the last one would likely be censored according Meta’s guidelines of supporting terrorism. In reality, Meta censors them all equally. The hashtags above are censored not because of what they stand for, but because the content enforcement models flagged too many posts that fall under these hashtags.

#istandwithhamas didn’t get censored for failing Meta’s guidelines for supporting Hamas.

#israel_hamas_are_murders didn’t get censored because Meta has a pro-Palestinian bias.

#sanctionisraelnow didn’t get censored because Meta has a pro-Israel bias.

Everyday, these hashtags are either removed or allowed to remain by the content enforcement models. We have even seen instances where one day these hashtags are blocked, and on another day they are visible. This is even acknowledged by Meta internally.

One of the comments I have heard for the Meta leaks part 1, was that I didn't show individual comments that were wrongly removed for their pro-Palestinian content. To which I argue, that anyone familiar with how ML models work at scale, knows that looking at trends is much more important than looking at individual posts (besides Human Right Watch already has data on individual posts). This is because when you deal with complex systems, probabilities, and billions of posts, you are likely to find examples that are for and against the argument you are trying to present. Hashtags are the highest level interpretable natural language unit representing topics and sentiment on social media. And in the case of this story, pro-Palestinian hashtags and posts are likely to get censored on Meta's platform.

Case 3: Bangladesh

The Bangladesh student led anti-government protests was the largest domestic political events for Bangladesh in 2024.

Coincidentally, the Bengali government is one of the biggest users of Meta's government initiated takedown requests (as presented in 'The Meta leaks part 1'). The government of Bangladesh is responsible for the sixth highest number of reports on content on Meta's platform.

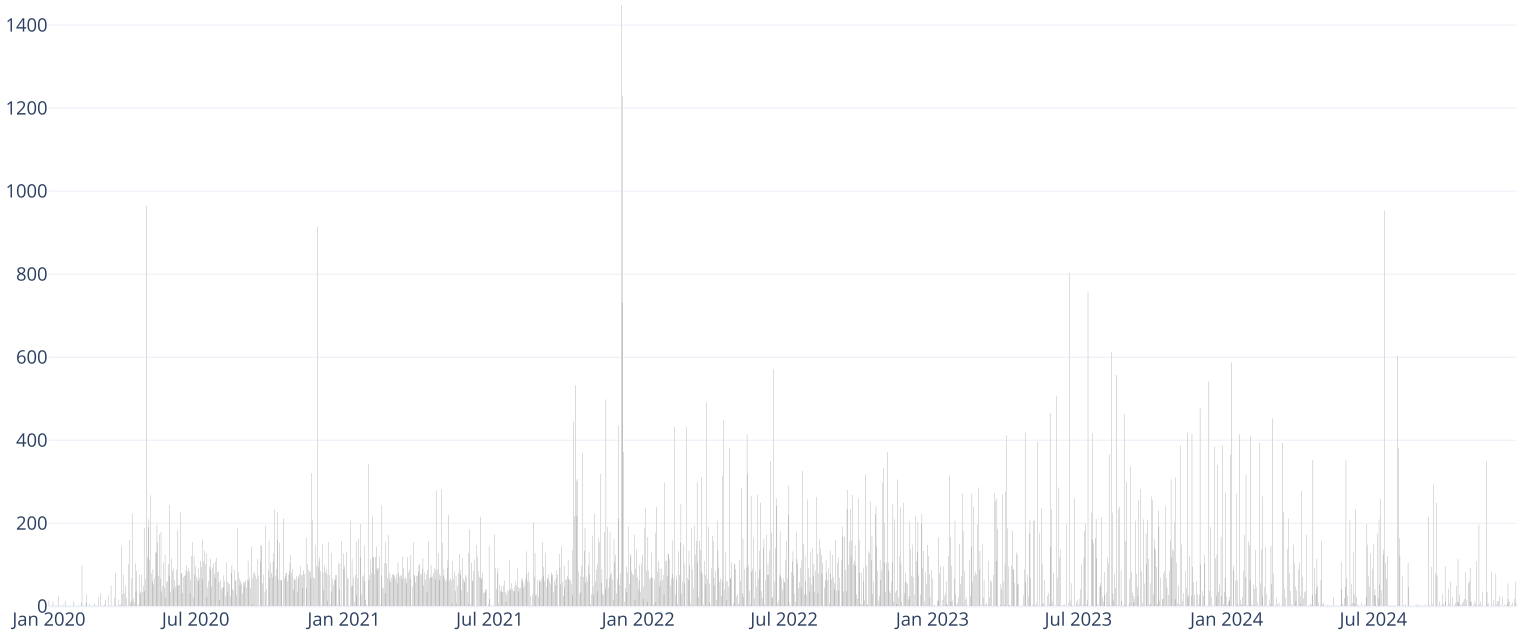

Figure 7. Number of posts targeted by takedown requests from the government of Bangladesh since 2024

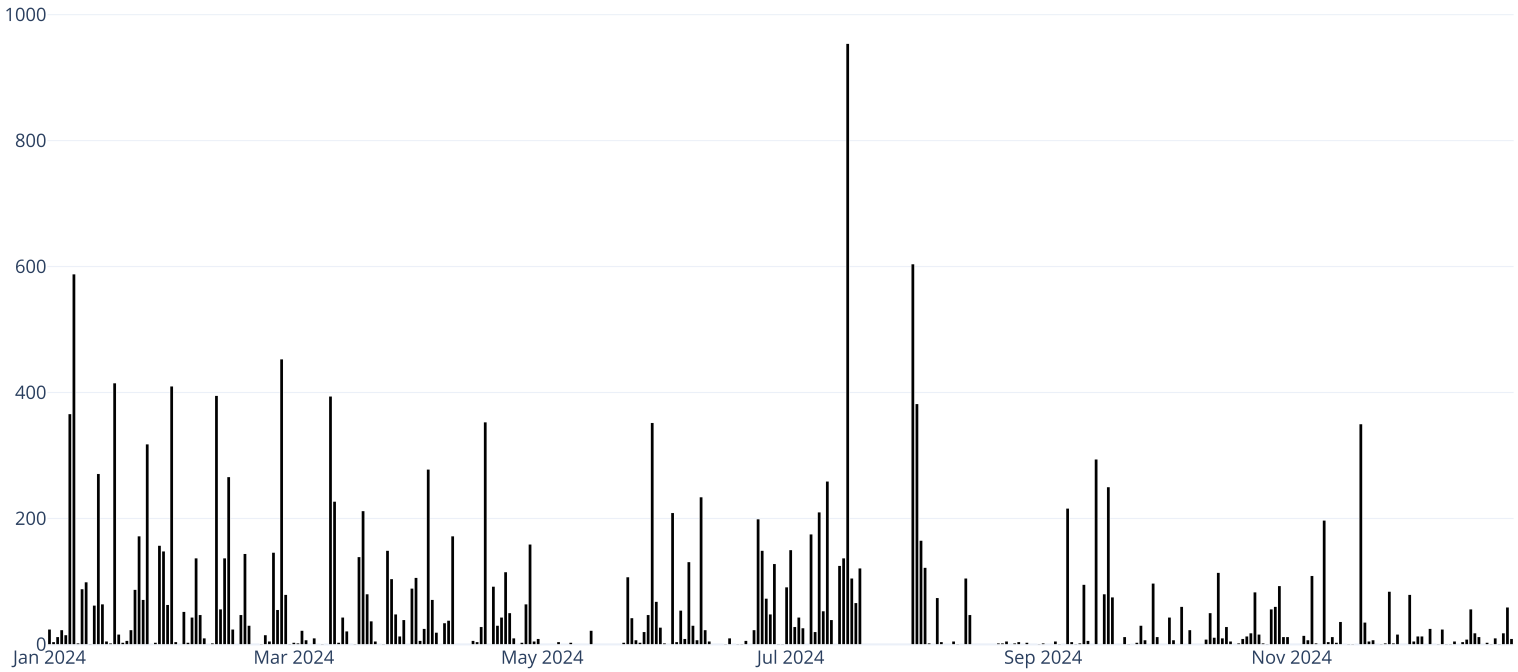

Figure 8. Number of posts targeted by takedown requests from the government of Bangladesh in 2024

If we look at the number of posts targeted in 2024, we can see how spikes in the data correspond with significant events related to the student protests of 2024. The number of posts targeted significantly decreases once the ruling government is removed from power.

The top three days with the most takedown requests are associated with major events/stories.

Date	# Reported posts	Event
2024-07-15	954	At least 100 wounded as Bangladesh students protest government job quotas (Al Jazeera)
2024-07-31	604	Bangladesh mourns some 200 deaths as student protests wind down and thousands are arrested (AP)
2024-01-07	588	Bangladesh Prime Minister Sheikh Hasina secures fourth term in election boycotted by opposition (CNN)

Discussion – Meta is complicit

Meta has a history of putting profits above people. There are countless examples of this, from them promoting violence against minority groups in Myanmar, their secret collaboration with the Israeli government in censoring pro-Palestinian and anti-Israel content, the way they treat their poorly paid outsourced moderators, and much more.

We also put forth the following claim against Meta. Meta is complicit in the mass censorship of posts and information that are orchestrated by bad actors (specifically oppressive governments). This is because:

- Meta is aware of the fact that their content enforcement models are flawed
- Meta has not taken any concrete steps to provide long-term fixes to their system
- Any fix done by Meta so far has been short-term (e.g. whitelisting false-positives hashtags)
- Abuses of Meta's platforms can lead to real life consequences (as in the case of Myanmar and Israel/Palestine)

There is overwhelming evidence from external and internal reports that Meta is aware of all the injustices done using their platform. This investigation sheds light on just some of them.